

MHC Class II Binding Prediction by Molecular Docking

M. Atanasova,^{*[a]} I. Dimitrov,^[a] D. R. Flower,^[b] and I. Doytchinova^[a]

Presented at the 18th European Symposium on Quantitative Structure Activity Relationships, EuroQSAR 2010, Rhodes, Greece

Abstract: Proteins of the Major Histocompatibility Complex (MHC) bind self and nonself peptide antigens or epitopes within the cell and present them at the cell surface for recognition by T cells. All T-cell epitopes are MHC binders but not all MHC binders are T-cell epitopes. The MHC class II proteins are extremely polymorphic. Polymorphic residues cluster in the peptide-binding region and largely determine the MHC's peptide selectivity. The peptide binding site on MHC class II proteins consist of five binding pockets. Using molecular docking, we have modelled the interactions between peptide and MHC class II proteins from locus DRB1.

Keywords: MHC class II · MHC-binding · Epitopes · Docking · Immunology

A combinatorial peptide library was generated by mutation of residues at peptide positions which correspond to binding pockets (so called anchor positions). The binding affinities were assessed using different scoring functions. The normalized scoring functions for each amino acid at each anchor position were used to construct quantitative matrices (QM) for MHC class II binding prediction. Models were validated by external test sets comprising 4540 known binders. Eighty percent of the known binders are identified in the best predicted 15% of all overlapping peptides, originating from one protein.

1 Introduction

T-Cell recognition is a fundamental mechanism underlying the cellular adaptive immune system by which the host identifies and responds to foreign antigens.^[1] The T cell is a specialized type of immune cell mediating cellular immunity. The T-cell receptors or TCRs, found on the surface of T cells, bind major histocompatibility complex (MHC) proteins, which are presented on antigen-presenting cell (APCs) surfaces. T-Cell immune responses are driven by recognition of peptide antigens (T-cell epitopes) that are bound to MHC molecules. Human MHC proteins, also known as human leukocyte antigens (HLA), are glycoproteins, which bind small peptide fragments, or epitopes, derived from both pathogen and host protein and present them at the cell surface for recognition by T cells.

There are two classes of MHC molecules: class I and class II. MHC class I molecules typically present peptides from proteins synthesized within the cell (endogenous processing pathway). MHC class II proteins primarily present peptides derived from endocytosed extracellular proteins (exogenous processing pathway).

Most nucleated cells express class I MHC proteins; these are recognized by T cells which highly express CD8 coreceptors. Class II MHCs are only expressed by so-called "professional antigen presenting cells" and are recognized by T cells which highly express CD4 coreceptors.

Both classes of MHC proteins are extremely polymorphic. More than 3500 molecules are listed in IMGT/HLA database.^[2] MHC class I proteins are encoded by three loci: HLA-A, HLA-B and HLA-C. MHC class II proteins are also encoded by three loci: HLA-DR, HLA-DQ and HLA-DP. The

peptide binding site of class I proteins has a closed cleft, formed by a single protein chain (α -chain).^[3] Usually, but not exclusively, only short peptides of 8–11 amino acids bind in extended conformation. In contrast, the cleft of class II proteins is open-ended, allowing much longer peptides to bind, although only 9 amino acids actually occupy the site. The class II cleft is formed by two separate protein chains: α and β .^[3] Clefs in class I and class II MHCs have binding pockets, corresponding to primary and secondary anchor positions on the binding peptide. The combination of two or more anchors is designated a motif. The experimental determination of motifs for every allele is prohibitively expensive in terms of labor, time and resources. The only practical alternative is to use a bioinformatics approach.

Different computational techniques have been applied to predict MHC class II binding. Most of them are sequence-based approaches such as matrix models,^[4–6] Artificial Neural Network,^[7–10] Hidden Markov Models,^[11] Support Vector Machines^[12,13] and those based on QSAR.^[14] Sophisticated methods, such as an iterative "meta-search" algo-

[a] M. Atanasova, I. Dimitrov, I. Doytchinova
Faculty of Pharmacy, Medical University of Sofia
2 Dunav str, 1000 Sofia, Bulgaria
phone: + 359 2 9236599; fax: + 359 2 9879874
*e-mail: matanasova@pharmfac.acad.bg

[b] D. R. Flower
Life and Health Sciences, Aston University
Aston Triangle, Birmingham, B4 7ET, UK

rithm^[15] and Ant Colony search^[16] have been created to resolve the dynamic variable-length problem inherent within the class II prediction. Certain new approaches have significantly outperformed more traditional methods.^[17,18]

The aim of the present study is to derive quantitative matrices (QMs) for MHC class II binding prediction by utilizing the technique of structure-based molecular docking. A combinatorial library of 95 peptides (19 aa × 5 pockets) was docked into the binding sites of 12 HLA-DRB1 proteins. Docking scores were normalized to assess the contribution of each amino acid at each pocket. Pocket polymorphism was described by different pocket profiles, with each HLA-DRB1 protein represented as a combination of five pocket profiles (pocket names correspond to peptide primary and secondary anchor positions). The specific docking score-based quantitative matrices (DS-QMs) obtained were tested on external dataset of 4540 known HLA-DRB1 binders.

2 Computational Methods

2.1 Molecular Modelling

Three X-ray structures of peptide-HLA-DRB1 protein complexes were used as input data for the molecular modelling in the present study. These were 1j8h (DRB1*0401, resolution 2.40 Å),^[19] 2g9h (DRB1*0101, resolution 2.0 Å)^[20] (Figure 1) and 1sje (DRB1*0101, resolution 2.45 Å).^[21]

HLA-DRB1 binding site consists of five pockets named after the corresponding peptide primary and secondary anchor positions: 1, 4, 6, 7 and 9. A set of profiles for each pocket was modelled by homology. HLA-DRB1 proteins

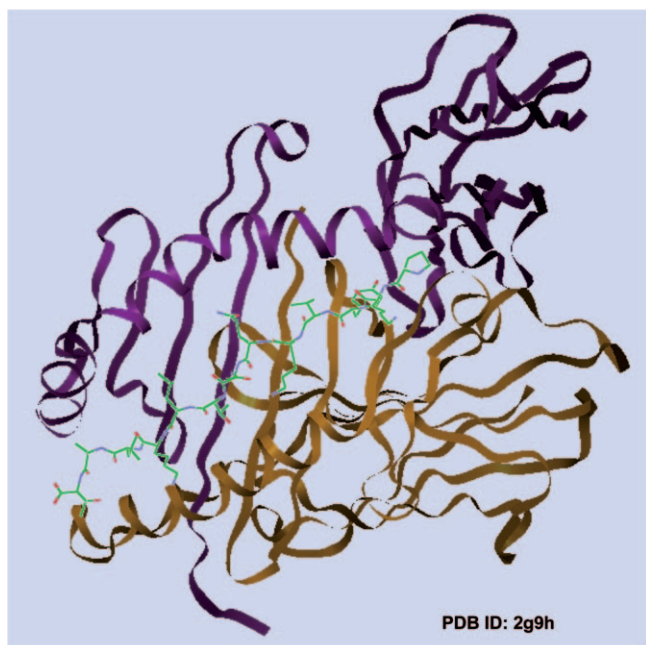


Figure 1. Crystal structure of staphylococcal enterotoxin I in complex with HLA-DR*0101 (pdb code: 2g9h).^[20]

were presented as combinations of five pocket profiles. Twelve HLA-DRB1 proteins were considered in the present study: DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0404, DRB1*0405, DRB1*0701, DRB1*0802, DRB1*0901, DRB1*1101, DRB1*1201, DRB1*1302 and DRB1*1501.

A library of 95 peptides (19 aa for 5 binding pockets) was generated for each HLA-DRB1 protein. The original bound peptides from the X-ray peptide-protein complexes were used as starting conformations.

All structures used in this study were modelled using PyMol v 0.99.^[22]

2.2 Docking Protocol

GOLD v. 4.1.2^[23] was used in the present study for molecular docking. The binding site was defined within 5 Å radius from the C α of the tested peptide position. The best ranking pose (lowest RMS) for each docking was chosen. Both protein and peptide were fixed apart from the tested peptide position and the corresponding binding pocket residues. The docking procedure was calibrated in regard to four criteria: 1) rigid versus flexible peptide position; 2) rigid versus flexible binding pocket residues; 3) ChemScore versus GoldScore scoring function; and 4) hard (6–12) versus soft (8–4) potential for binding pocket residues. Thus 16 different parameter combinations were evaluated. The average score of five runs for each docking was taken. The docking scores of all 20 amino acids for each pocket were normalized to derive the pocket-specific amino acid contributions. Thresholds were defined to distinguish the most preferred and the most deleterious amino acids: > 0.3 for the most preferred and < -0.2 for the most deleterious.

Each of the 12 HLA-DRB1 proteins was presented as a combination of five pocket profiles and acquired a specific docking score-based quantitative matrix (DS-QM).

2.3 External Validation

The predictive ability of DS-QMs was tested using a set of 4540 known peptide binders originating from 167 proteins. The peptides were of different length and bound to the studied 12 HLA-DRB1 proteins. Data were extracted from the Immune Epitope Database (<http://www.immuneepitope.org>)^[24] in December 2009, according to the following criteria: Alleles: DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0404, DRB1*0405, DRB1*0701, DRB1*0802, DRB1*0901, DRB1*1101, DRB1*1201, DRB1*1302 and DRB1*1501; Assay: Purified MHC – Radioactivity Competition; Quantitative measurement; Units: IC₅₀ nM. From a set of overlapping peptides with different lengths, only the longest peptide was included in the test set. Peptides binding affinities were originally assessed using a quantitative assay based on the inhibition of binding of a radio-labelled standard peptide to detergent-solubilized MHC molecules^[25,26] and presented as $pIC_{50} = \log(1/IC_{50})$.

Each protein of origin was presented as a set of overlapping nonamers and the binding score of each nonamer was predicted by the derived DS-QMs. The predicted p/C_{50} values were arranged in a descending mode and only the best predicted 5%, 10% and 15% of all nonamers originating from one protein were selected and compared to the known binders from the same protein. A binder is considered as a true predicted binder, if it contains any nonamer sequence from the predicted top sets. The total sum of true predicted binders was used to assess the predictive ability via the parameter *sensitivity*:

$$\text{sensitivity} = 100 \times \text{true predicted binders/all binders}$$

3 Results

Three libraries of 19 peptides were modeled for each of the three X-ray peptide-protein complexes used (pdb codes: 1j8h, 2gh and 1sje). The corresponding bound peptides for each structure were used as initial conformations. Substitutions were made at anchor position 1. Peptides were docked applying the docking protocol described in Computational methods. Similar binding scores were generated for the three complexes (data not shown). Therefore, subsequently only one X-ray structure was used: pdb code 2g9h.

Lipophilic amino acids at peptide position 1, as Phe, Tyr, Trp, Leu, Ile, Met and Val are preferred for binding to HLA-DRB1 proteins.^[27] The docking procedure was calibrated to fulfill these preferences. The best agreement was achieved using the following parameter combination: 1) flexible peptide position; 2) rigid binding pocket residues; 3) ChemScore scoring function; and 4) soft (8–4) potential for binding pocket residues. These conditions were used in subsequent docking procedures for all pockets.

All docking solutions in the present study had RMS values below or close to 1Å.

3.1 Docking Scores for Pocket Profiles

Pocket 1 profiles. A dimorphism in pocket 1 exists for HLA-DRB1 proteins. Proteins DRB1*0101, DRB1*0401, DRB1*0405, DRB1*0701, DRB1*0802, DRB1*0901, DRB1*1101, and DRB1*1301 possess Gly at position 86 β , while DRB1*0301, DRB1*0404, DRB1*1201, and DRB1*1501 contain Val at the same position. Pockets with Gly^{86 β} were defined as pocket profile 1A (Figure 2A), while pockets with Val^{86 β} – as pocket profile 1B (Figure 2B).

Table 1 summarises the most preferred and the most deleterious amino acids for profiles 1A and 1B. It is evident that the mutation Gly^{86 β} →Val^{86 β} makes the pocket shallow and rejects Tyr as preferred amino acid.

Pocket 4 profiles. Pocket 4 is polymorphic in positions 13 β , 70 β , 71 β , 74 β and 78 β (Figure 2C). Eleven different profiles could be distinguished here among the studied 12 HLA-DRB1 proteins.

The preferred and deleterious amino acids for the different pocket 4 profiles are given in Table 1.

Position 74 β is primarily responsible for pocket 4 preferences.^[28,29] If short side-chain amino acids, such as Ala, Gln and Glu, occupy this position, the pocket could accept peptides with long chain amino acids at p4, like Tyr, Trp and Phe. Otherwise, Lys and Arg at 74 β shorten the pocket and only small chain amino acids could be placed here.

Pro is the common deleterious amino acid for all pocket 4 profiles. Additionally, when Lys or Arg is present at position 71 β , Arg is the second common deleterious amino acid. Mutations to Glu or Ala reject Arg from the nonpreferred amino acid list.

Pocket 6 profiles. Pocket 6 contains only one polymorphic position, namely 11 β (Figure 2D).

The preferred and deleterious amino acids for pocket 6 profiles are given in Table 1.

Generally, pocket 6 is shallow but wide. For that reason, Tyr is deleterious here (apart from profile 6D which contains Gly), but Trp is well accepted. Medium sized amino acids, like Met, Leu and Ile, fit well in this pocket.

Pocket 7 profiles. Pocket 7 contain five polymorphic positions: 28 β , 30 β , 47 β , 67 β and 71 β (Figure 2E). They define 11 profiles for the studied HLA-DRB1 proteins. The preferred and deleterious amino acids for pocket 7 profiles are given in Table 1.

Similarly to pocket 6, pocket 7 is shallow and wide. It accepts well Trp and Phe. Pro is also welcome here, and this could be due to adopting a more preferred conformation rather than by a size preference.

The most deleterious amino acid here is Arg because of the electrostatic repulsion with Arg/Lys^{71 β} .

Pocket 9 profiles. Four polymorphic amino acids in pocket 9 generate eight profiles. These are positions 9 β , 37 β , 57 β and 60 β (Figure 2F).

Table 1 summarises the top preferred and deleterious amino acids for the pocket 9 profiles.

The polymorphism at 9 β determines the pocket size and hence the preferred amino acids. Phe is the preferred residue here. For profiles containing the negatively charged Glu^{9 β} , Lys is also a preferred amino acid. Pro and Arg are common deleterious amino acids for pocket 9.

3.2 Docking Score-Based Quantitative Matrixes (DS-QM)

The HLA-DRB1 proteins considered in the present study were as follows: DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0404, DRB1*0405, DRB1*0701, DRB1*0802, DRB1*0901, DRB1*1101, DRB1*1201, DRB1*1302 and DRB1*1501. Each protein was presented as a combination of five pocket profiles (Table 2) and acquired a specific docking score-based quantitative matrix (DS-QM). The DS-QMs for the 12 HLA-DRB1 proteins are given as Supporting Information at www.molinf.com. Amino acids with positive coefficients increase affinity of peptides for HLA-DRB1 proteins, those with negative coefficients decrease it.

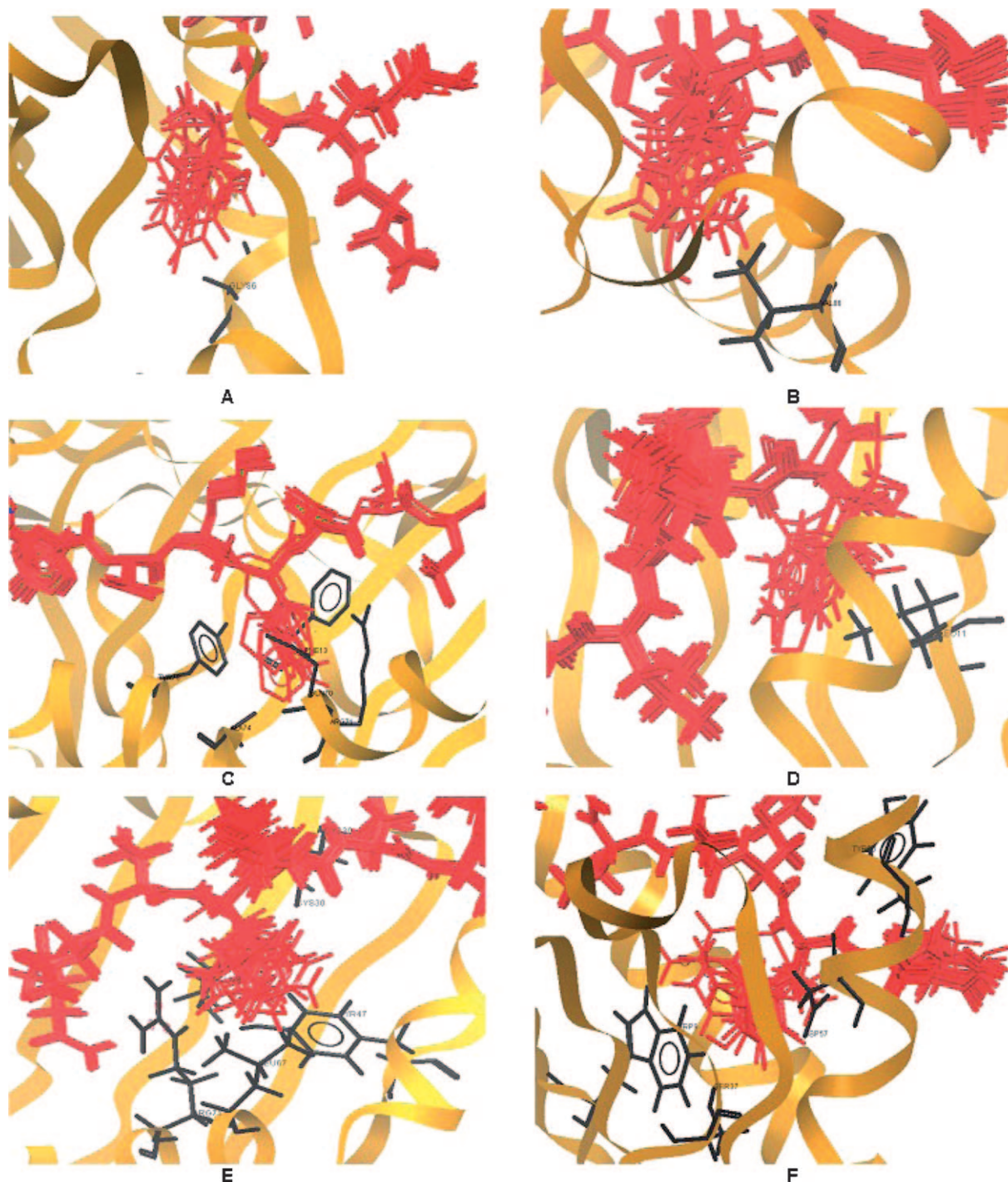


Figure 2. Binding pocket profiles: A) 1A; B) 1B; C) 4A; D) 6A; E) 7A; F) 9A. Only the polymorphic amino acid is given (in black). The best docking poses of a library of 20 peptides (1 original + 19 modelled) are shown aligned (in dark grey).

3.3 Prediction by DS-QM

The predictive ability of the DS-QMs was assessed by an external test set of 4540 known peptide binders to the 12 studied HLA-DRB1 proteins. The binders originate from 167 proteins. Each protein was represented as a set of overlapping nonamer peptides and the binding affinity of each

peptide to each HLA-DRB1 protein was predicted by the corresponding DS-QM. Peptides were ranked in descending order according to their predicted binding score. The top 5% were selected and compared to the known binders for this DRB1 protein. If the predicted nonamer was contained in the sequence of the known binder, the prediction is con-

Table 1. Preferred and deleterious amino acids residues from peptides for defined protein pocket profiles.

| Pocket ID | Protein pocket profile | Peptide | |
|-----------|--------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------|
| | | Preferred aa Score > 0.3 | Deleterious aa Score < -0.2 |
| 1A | Gly ^{86β} | Trp, Phe, Tyr | Pro, Asp, Glu, Gly |
| 1B | Val ^{86β} | Phe, Trp | Pro, Arg, Asp |
| 4A | Phe ^{13β} , Gln ^{70β} , Arg ^{71β} , Ala ^{74β} , Tyr ^{78β} | Phe, Lys, ^[a] Leu, ^[a] Met ^[a] | Pro, Arg, Asn |
| 4B | Ser ^{13β} , Gln ^{70β} , Lys ^{71β} , Arg ^{74β} , Tyr ^{78β} | Ser, Ala, ^[a] Leu ^[b] | Pro |
| 4C | His ^{13β} , Gln ^{70β} , Lys ^{71β} , Ala ^{74β} , Tyr ^{78β} | Trp, Leu | Pro, Arg |
| 4D | His ^{13β} , Gln ^{70β} , Arg ^{71β} , Ala ^{74β} , Tyr ^{78β} | Trp, Phe | Pro, Gly, Asn, Arg, Asp, Ser |
| 4E | Tyr ^{13β} , Asp ^{70β} , Arg ^{71β} , Gln ^{74β} , Val ^{78β} | Ala, ^[a] Cys, ^[b] Gly ^[b] | Trp, Gln, Asp, Glu |
| 4F | Gly ^{13β} , Asp ^{70β} , Arg ^{71β} , Lys ^{74β} , Tyr ^{78β} | Met, Val, Leu | Pro, His, Asn, Gln, Ala, Arg |
| 4G | Phe ^{13β} , Arg ^{70β} , Arg ^{71β} , Glu ^{74β} , Val ^{78β} | Ser, Met, Val, Phe | Gln, Asn, Pro, Asp, Gly, Arg |
| 4H | Ser ^{13β} , Asp ^{70β} , Arg ^{71β} , Ala ^{74β} , Tyr ^{78β} | Asp, ^[b] Ile, ^[b] Leu ^[b] | Pro, Arg |
| 4I | Gly ^{13β} , Asp ^{70β} , Arg ^{71β} , Ala ^{74β} , Tyr ^{78β} | Leu, Phe, Ile | Pro, Glu, Arg, Cys |
| 4J | Ser ^{13β} , Asp ^{70β} , Glu ^{71β} , Ala ^{74β} , Tyr ^{78β} | Trp, Phe | Pro, Gly, Asp |
| 4K | Arg ^{13β} , Gln ^{70β} , Ala ^{71β} , Ala ^{74β} , Tyr ^{78β} | Gly, Ala, Tyr | Pro, Val, Gln, Asn, Asp |
| 6A | Leu ^{11β} | Glu, Met ^[a] | Tyr, Asp, Phe |
| 6B | Ser ^{11β} | Ser, Ala, ^[a] Trp ^[a] | Tyr, Pro, Leu, Phe |
| 6C | Val ^{11β} | Met, ^[a] Thr, ^[a] Val, ^[a] Trp ^[a] | Tyr, Pro, Asp, Leu |
| 6D | Gly ^{11β} | Trp, Phe | Asp, Pro, Leu |
| 6E | Asp ^{11β} | Ala, ^[a] Glu, ^[a] Ile ^[a] | Tyr, Asp, Leu, Phe |
| 6F | Pro ^{11β} | Met, Trp | Pro, Asp, Tyr, Leu, His, Asn |
| 7A | Glu ^{28β} , Cys ^{30β} , Tyr ^{47β} , Leu ^{67β} , Arg ^{71β} | Pro, Phe, ^[a] Ile ^[a] | Asp, Arg, Gln |
| 7B | Asp ^{28β} , Tyr ^{30β} , Phe ^{47β} , Leu ^{67β} , Lys ^{71β} | Leu, Trp, Phe | Arg, His, Ser, Gln |
| 7C | Asp ^{28β} , Tyr ^{30β} , Tyr ^{47β} , Leu ^{67β} , Lys ^{71β} | Phe, Trp | Arg, Ser, Asp, Gln, His |
| 7D | Asp ^{28β} , Tyr ^{30β} , Tyr ^{47β} , Leu ^{67β} , Arg ^{71β} | Trp, Phe, Leu | Glu, Arg, Gln, Ser |
| 7E | Glu ^{28β} , Leu ^{30β} , Tyr ^{47β} , Ile ^{67β} , Arg ^{71β} | Phe, Pro, Tyr, Trp | Arg, Gln, Ser, Asp, Thr |
| 7F | Asp ^{28β} , Tyr ^{30β} , Tyr ^{47β} , Phe ^{67β} , Arg ^{71β} | Trp, Phe, Met | Arg, Glu, Gln, Asn, His |
| 7G | His ^{28β} , Gly ^{30β} , Tyr ^{47β} , Phe ^{67β} , Arg ^{71β} | Phe, Pro, Trp | Asn, Arg, Ser, Gln, Asp, Glu |
| 7H | Asp ^{28β} , Tyr ^{30β} , Phe ^{47β} , Phe ^{67β} , Arg ^{71β} | Trp, Pro, Phe | Ser, Asn, Gln, Glu, Asp, Arg |
| 7I | Glu ^{28β} , His ^{30β} , Phe ^{47β} , Ile ^{67β} , Arg ^{71β} | Trp, Pro | Arg, Gln, Glu, His, Asp, Asn |
| 7J | Asp ^{28β} , Tyr ^{30β} , Phe ^{47β} , Ile ^{67β} , Glu ^{71β} | Trp, Phe | Asn, Asp, Ser, Gln, Glu, Thr |
| 7K | Asp ^{28β} , Tyr ^{30β} , Phe ^{47β} , Ile ^{67β} , Ala ^{71β} | Trp, Phe | Asn, Glu, Arg, Gln, Thr, Asp |
| 9A | Trp ^{9β} , Ser ^{37β} , Asp ^{57β} , Tyr ^{60β} | Leu, Phe, Met ^[a] | Pro, Arg |
| 9B | Glu ^{9β} , Asn ^{37β} , Asp ^{57β} , Tyr ^{60β} | Phe, Tyr | Pro |
| 9C | Glu ^{9β} , Tyr ^{37β} , Asp ^{57β} , Tyr ^{60β} | Lys, Phe, Tyr ^[a] | Pro, Arg |
| 9D | Glu ^{9β} , Tyr ^{37β} , Ser ^{57β} , Tyr ^{60β} | Lys, Phe | Pro, Trp, Arg |
| 9E | Trp ^{9β} , Phe ^{37β} , Val ^{57β} , Ser ^{60β} | Phe, Leu ^[a] | Pro, Arg |
| 9F | Glu ^{9β} , Tyr ^{37β} , Asp ^{57β} , Tyr ^{60β} | Lys, Phe, Leu ^[a] | Pro, Arg |
| 9G | Lys ^{9β} , Asn ^{37β} , Val ^{57β} , Ser ^{60β} | Phe, Tyr, Trp ^[a] | Pro, Arg |
| 9H | Glu ^{9β} , Leu ^{37β} , Val ^{57β} , Ser ^{60β} | Phe, Tyr | Pro |

[a] score > 0.25; [b] score > 0.2.

Table 2. Pockets profiles corresponding to 12 HLA-DR alleles are presented.

| Allele | Pocket profiles | | | | |
|-----------|-----------------|----|----|----|----|
| DRB1*0101 | 1A | 4A | 6A | 7A | 9A |
| DRB1*0301 | 1B | 4B | 6B | 7B | 9B |
| DRB1*0401 | 1A | 4C | 6C | 7C | 9C |
| DRB1*0404 | 1B | 4D | 6C | 7D | 9C |
| DRB1*0405 | 1A | 4D | 6C | 7D | 9D |
| DRB1*0701 | 1A | 4E | 6D | 7E | 9E |
| DRB1*0802 | 1A | 4F | 6B | 7F | 9F |
| DRB1*0901 | 1A | 4G | 6E | 7G | 9G |
| DRB1*1101 | 1A | 4H | 6B | 7H | 9C |
| DRB1*1201 | 1B | 4I | 6B | 7I | 9H |
| DRB1*1302 | 1A | 4J | 6B | 7J | 9B |
| DRB1*1501 | 1B | 4K | 6F | 7K | 9A |

sidered a “true predicted binder”. The ratio of true predicted binders to all known binders gives the *sensitivity* of prediction. The same procedure was applied to the top 10% and 15% of the predicted nonamers.

The sensitivity was assessed “pocket by pocket” and “all pockets”. Average values of all 12 HLA-DRB1 proteins are given at Figure 3. The “pocket by pocket” sensitivity is around 30% for the top 5% of the predicted binders, be-

tween 50% and 60% – for the top 10%, and above 70% for the top 15%. The “all pockets” sensitivity is slightly better than that of single pockets. It starts from 40% for the top 5% and reaches 80% for the top 15%. No additivity exists between pockets. This result confirms once again that the binding on MHC is not additive and rejects the hypothesis of Independent Binding of Side chains (IBS hypothesis).^[30]

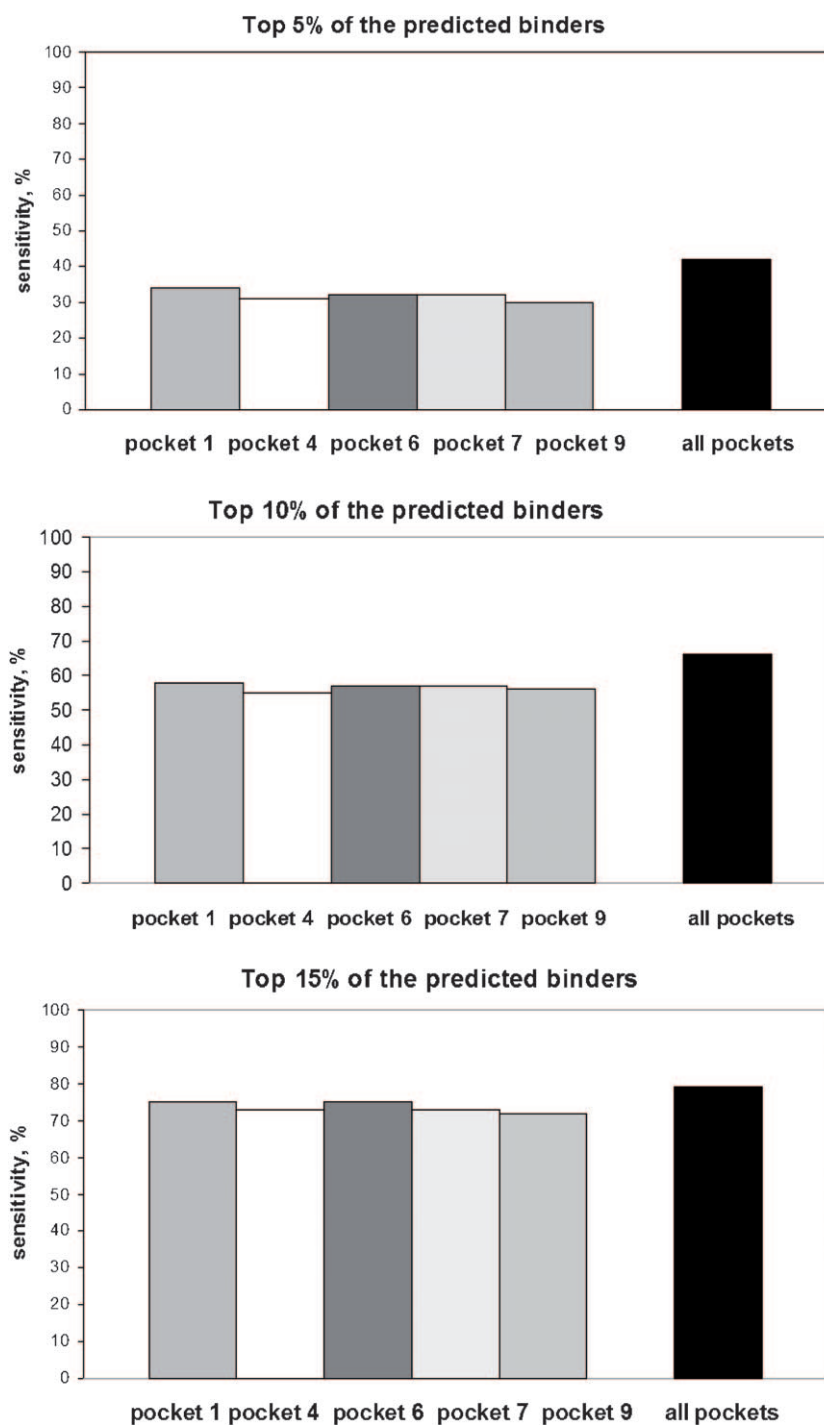


Figure 3. Average sensitivity for single pockets and all pockets at top 5%, 10% and 15% of the best predicted binders.

4 Discussion

The main prerequisite for a peptide to act as a T-cell epitope is that it binds to an MHC protein. Because of the large number of MHC proteins and antigenic peptides, the experimental identification of MHC binders is a prohibitively time consuming and expensive process. The only tractable alternative approach is MHC binding prediction using computational methods.

There is a great variety of computational methods for MHC binding prediction. They can be classified into two groups: sequence-based and structure-based. Sequence-based approaches develop prediction models based on large datasets of experimental data. Such methods are motif searching,^[31] quantitative matrices,^[4–6,32,33] Artificial Neural Networks,^[7–10,34–36] Hidden Markov Models,^[11,37] and Support Vector Machines.^[12,13,38–40] Structure-based methods make use of the 3D structure of the binding peptide – MHC protein interface and analyze the interactions between them. Here are included protein threading,^[41,42] homology modelling,^[43,44] and docking.^[45–47] A significant potential advantage of structure-based methods is their ability to predict binding to structures where experimental data are absent or are insufficient. The aim of both ligand-based and structure-based MHC binding prediction is to identify viable biophores that interact with the great variety of binding sites implicit within the population of MHC molecules. In immunology, such biophores are typically termed motifs.

The method described in the present study is a structure-based one utilizing the technique of molecular docking to derive quantitative models for MHC class II binding prediction. The only input data used in the models development are the X-ray structures of known peptide – MHC protein complexes. The method was applied to 12 HLA-DRB1 proteins and tested on an external set of 4540 known binders. It recognizes 80% of the true binders in the best predicted 15% of all overlapping peptides, originating from one protein. Due to the high resource implications of experimental testing, when scanning a large proteome high numbers of false positives present a greater problem than high numbers of false negatives. Taking into account only the best predicted binders significantly reduces the number of false positives.

ChemScore is an empirical function which uses four terms to approximate individual contributions to the binding energy: hydrogen bonding, lipophilic interactions, metal-ligand binding and loss of ligand flexibility.^[48] The ChemScore implemented in GOLD contains additional terms accounting for atom-atom clashes and internal torsion interactions, which compensate for close contacts and poor internal conformations encountered during docking.^[49] Original ChemScore uses block functions. The GOLD implementation of ChemScore, uses a Gaussian-smoothed function to account for contact terms. The hydrogen-bond term is a sum over all possible donor-acceptor pairs, where

one atom belongs to the protein and the other to the ligand. The lipophilic term is the sum over all lipophilic atoms in the protein and the ligand. The metal-binding term is computed as a sum over all possible metal-ion acceptor pairs, where the acceptor is an atom in the ligand capable of binding to a metal. The rotatable-bond freezing term, used to estimate the entropic loss that occurs when single, acyclic bonds in the ligand become nonrotatable, accounts for frozen rotatable bonds in the ligand.

The term “pocket profile” was coined for the first time by Sturmiolo.^[4] Sturmiolo’s method based on MHC pocket profiles is an example of a structure-based method for MHC binding prediction. Each MHC pocket on the binding site is determined by a set of amino acids; some are conserved, others are polymorphic. The interactions made by all natural amino acids with a given pocket establish the pocket profile. Pocket profiles are nearly independent of the remaining MHC binding site.

Peptides which bind to MHC proteins are extremely flexible molecules with very many low-energy conformations. Also, the binding site on class II proteins is open-ended which potentially allows a peptide to bind in several different registers.^[50] To remove a part of this uncertainty, many MHC class II predictors use the “one binder – one pattern” assumption. Another approximation used by QM methods is the additivity concept based on the IBS hypothesis,^[30] which considers the binding affinity as a linear sum of the binding affinities at each peptide position. However, peptide binding to MHC molecules is neither single pattern-based, position-independent, or linear additive. The absence of additivity was demonstrated in the present study.

In conclusion, the new structure-based method for MHC binding prediction described in the present study is a reliable tool for the large-scale screening of potential T-cell epitopes. It facilitates experimental laboratory work reducing significantly resource requirements, including practitioner times and labour.

Acknowledgements

The present research was supported by *The National Research Fund of Ministry of Education and Science*, Bulgaria (Grant 02-1/2009).

References

- [1] D. R. Flower, in *Bioinformatics for Vaccinology* (Ed: D. R. Flower), Wiley-Blackwell, Chichester, UK, **2008**, pp. 94–95.
- [2] J. Robinson, M. J. Waller, P. Parham, N. de Groot, R. Bontrop, L. J. Kennedy, P. Stoehr, S. G. E. Marsh, *Nucleic Acids Res.* **2003**, *31*, 311–314.
- [3] C. A. Janeway, P. Travers, M. Walport, J. D. Capra, in *Immunobiology: the Immune System in Health and Disease* (Ed: C. A. Janeway), Current Biology Publications, London, **1999**, pp. 115–162.

- [4] T. Sturniolo, E. Bono, J. Ding, L. Radrizzani, O. Tuereci, U. Sahin, M. Braxenthaler, F. Gallazzi, M. P. Protti, F. Sinigaglia, J. Hammer, *Nature Biotechnol.* **1999**, *17*, 555–561.
- [5] H.-H. Bui, J. Sidney, B. Peters, M. Sathiamurthy, A. Sinichi, K. A. Purton, B. R. Mothe, F. V. Chisari, D. I. Watkins, A. Sette, *Immunogenetics* **2005**, *57*, 304–314.
- [6] M. Nielsen, C. Lundegaard, O. Lund, *BMC Bioinformatics*, **2007**, *8*, article 238.
- [7] V. Brusica, G. Rudy, M. Honeyman, J. Hammer, L. Harrison, *Bioinformatics* **1998**, *14*, 121–130.
- [8] K. Gulukota, C. DeLisi, *Methods Molec. Biol.* **2001**, *156*, 201–209.
- [9] H. Noguchi, T. Hanai, H. Honda, L. C. Harrison, T. Kobayashi, *J. Biosci. Bioeng.* **2001**, *92*, 227–231.
- [10] F. R. Burden, D. A. Winkler, *J. Mol. Graph. Model.* **2005**, *23*, 481–489.
- [11] H. Noguchi, R. Kato, T. Hanai, Y. Matsubara, H. Honda, V. Brusica, T. Kobayashi, *J. Biosci. Bioeng.* **2002**, *94*, 264–270.
- [12] J. Wan, W. Liu, Q. Xu, Y. Ren, D. R. Flower, T. Li, *BMC Bioinformatics*, **2006**, *7*, article 463.
- [13] M. Bhasin, G. P. Raghava, *Bioinformatics* **2004**, *20*, 421–423.
- [14] I. Dimitrov, P. Garnev, D. R. Flower, I. Doytchinova, *Eur. J. Med. Chem.* **2010**, *45*, 236–243.
- [15] R. R. Mallios, *Bioinformatics* **2001**, *17*, 942–948.
- [16] O. Karpenko, J. Shi, Y. Dai, *Artif. Intell. Med.* **2005**, *35*, 147–156.
- [17] J. Salomon, D. R. Flower, *BMC Bioinformatics* **2006**, *7*, article 501.
- [18] W. Zhang, J. Liu, Y. Niu, *Artif. Intell. Med.* **2010**, *50*, 127–132.
- [19] J. Hennecke, D. C. Wiley, *J. Exp. Med.* **2002**, *195*, 571–581.
- [20] M. M. Fernandez, R. Guan, C. P. Swaminathan, E. L. Malchiodi, R. A. Mariuzza, *J. Biol. Chem.* **2006**, *281*, 25356–25364.
- [21] Z. Zavala-Ruiz, I. Strug, B. D. Walker, P. J. Norris, L. J. Stern, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 13279–13284.
- [22] W. L. DeLano, *The Pymol Molecular Graphics System 2006*, DeLano Scientific, San Carlos, CA; <http://pymol.sourceforge.net/>
- [23] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727–748.
- [24] R. Vita, L. Zarebski, J. A. Greenbaum, H. Emami, I. Hoof, N. Salimi, R. Damle, A. Sette, B. Peters, *Nucleic Acids Res.* **2010**, *38* (Database issue), D854–862.
- [25] J. Ruppert, J. Sidney, E. Celis, R. T. Kubo, H. M. Grey, A. Sette, *Cell* **1993**, *74*, 929–937.
- [26] A. Sette, J. Sidney, M.-F. Del Guercio, S. Southwood, J. Ruppert, C. Dalberg, H. M. Grey, R. T. Kubo, *Mol. Immunol.* **1994**, *31*, 813–822.
- [27] H.-G. Rammensee, J. Bachmann, N. N. Emmerich, O. A. Bachor, S. Stevanovic, *Immunogenetics* **1999**, *50*, 213–219.
- [28] I. A. Doytchinova, D. R. Flower, *J. Immunol.* **2005**, *174*, 7085–7095.
- [29] D. Ou, L. A. Mitchell, A. J. Tingle, *Hum. Immunol.* **1998**, *59*, 665–676.
- [30] K. C. Parker, M. A. Bednarek, J. E. Coligan, *J. Immunol.* **1994**, *152*, 163–175.
- [31] M. Nielsen, C. Lundegaard, P. Worning, C. S. Hvid, K. Lamberth, S. Buus, S. Burak, O. Iund, *Bioinformatics* **2004**, *20*, 137–148.
- [32] M. Vordermeier, A. O. Whelan, R. G. Hewinson, *Infect. Immun.* **2003**, *71*, 1860–1867.
- [33] I. A. Doytchinova, D. R. Flower, *Mol. Immunol.* **2006**, *40*, 2037–2044.
- [34] C. Schonbach, Y. Kun, V. Brusica, *Immunol. Cell Biol.* **2002**, *80*, 300–306.
- [35] M. Bhasin, G. P. Raghava, *J. Biosci.* **2007**, *32*, 31–42.
- [36] G. L. Zhang, A. M. Khan, K. N. Srinivasan, A. Heiny, K. Lee, C. K. Kwok, J. T. August, V. Brusica, *BMC Bioinformatics* **2008**, *9* (Suppl 1), S19.
- [37] K. N. Srinivasan, G. L. Zhang, A. M. Khan, J. T. August, V. Brusica, *Bioinformatics* **2004**, *20*, 297–302.
- [38] P. Donnes, A. Elofsson, *BMC Bioinformatics* **2002**, *3*, article 25.
- [39] S. Li, X. Yao, H. Liu, J. Li, B. Fan, *Anal. Chim. Acta.* **2007**, *584*, 37–42.
- [40] W. Liu, J. Wan, X. Meng, D. R. Flower, T. Li, *Mol. Biol.* **2007**, *409*, 283–291.
- [41] S. P. Singh, B. N. Mishra, *Bioinformation* **2008**, *3*, 72–82.
- [42] B. Knapp, U. Omasits, S. Frantal, W. Schreiner, *J. Comput. Aided Mol. Des.* **2009**, *23*, 301–307.
- [43] A. Logean, D. Rognan, *J. Comput. Aided Mol. Des.* **2002**, *16*, 229–243.
- [44] A. Kosmopoulou, M. Vlassi, A. Stavrakoudis, C. Sakarellos, M. Sakarellos-Daitsiotis, *J. Comput. Chem.* **2006**, *27*, 1033–1044.
- [45] N. Sauton, D. Lagorce, B. O. Villoutreix, M. A. Miteva, *BMC Bioinformatics* **2008**, *9*, article 184.
- [46] J. C. Tong, T. W. Tan, S. Ranganathan, *Protein Sci.* **2004**, *13*, 2523–2532.
- [47] J. C. Tong, G. L. Zhang, T. W. Tan, J. T. August, V. Brusica, S. Ranganathan, *Bioinformatics* **2006**, *22*, 1232–1238.
- [48] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, R. P. Mee, *J. Comput. Aided Mol. Des.* **1997**, *11*, 425–445.
- [49] *GOLD User Guide & Tutorials* (<http://www.ccdc.cam.ac.uk/support/documentation>).
- [50] I. A. Doytchinova, D. R. Flower, *Bioinformatics* **2003**, *19*, 2263–2270.

Received: October 27, 2010

Accepted: December 1, 2010

Published online: March 31, 2011